An integrative statistical model for inferring strain admixture within clinical *Plasmodium falciparum* isolates

John D O'Brien[1,*], Zamin Iqbal[2], and Lucas Amenga-Etego[2,3]

[1]Department of Mathematics, Bowdoin College, Brunswick, Maine, USA
[2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[3]Navrongo Health Research Centre, Navrongo, Upper East Region, Ghana
[*]To whom correspondence should be addressed

ABSTRACT

Since the arrival of genetic typing methods in the late 1960's, researchers have puzzled at the clinical consequence of observed strain mixtures within clinical isolates of *Plasmodium falciparum*. We present a new statistical model that infers the number of strains present and the amount of admixture with the local population (panmixia) using whole-genome sequence data. The model provides a rigorous statistical approach to inferring these quantities as well as the proportions of the strains within each sample. Applied to 168 samples of whole-genome sequence data from northern Ghana, the model provides significantly improvement fit over models implementing simpler approaches to mixture for a large majority (129/168) of samples. We discuss the possible uses of this model as a window into within-host selection for clinical and epidemiological studies and outline possible means for experimental validation.

INTRODUCTION

The protozoan parasite *Plasmodium falciparum* (Pf) is the cause of the vast majority of fatal malaria cases, killing at least half a million people a year [16, 42, 48]. The parasite's ability to develop resistance to drugs and the rapid spread of that resistance across geographically-separated populations presents a constant threat to international control efforts [47, 28, 34]. While research has elucidated many genetic factors in resistance, much of genetic epidemiology of the parasite - including the effective recombination rate and the rate of gene flow across populations - is still unclear [39, 28, 37].

Since the late 1960's, researchers focused on the structure of Pf clinical infections have struggled to understand the implications of multiplicity of infection (MOI), where multiple strains appear to be present within a single patient's bloodstream [46, 25, 19, 7, 3]. While MOI-focused studies implicate increased or decreased levels of MOI with a range of conditions, including clinical severity [29], age-specific severity [17, 40, 8, 44], parasitemia levels during pregnancy [5], and other effects [4, 41, 33, 24], there is no broad consensus about MOI's role, if any, in controling the course of an infection. Still, a wide variety of studies and genetic assays – most typically by typing of the *mdr* gene – show MOI as a reliable feature of clinical Pf isolates [3].

The appearance of whole-genome sequencing (WGS) technologies applied to Pf extracted directly from infected patients' bloodstreams provides an unprecedented window into the structure of genetic mixture within samples [23, 2]. Initial work on understanding the structure of this mixture data shifted focus from estimating MOI to analysis based on inbreeding coefficients [3, 31, 45]. These metrics, a special type of F-statistic, provide an estimate of the departure of within-sample allele frequencies from those expected under a Hardy-Weinberg-type equilibrium with the nearby population. In this perspective, each patient's bloodstream is taken to be a subpopulation exhibiting a degree of admixture of all of the strains from the local environment, ranging from a perfectly random sampling of all nearby strains to the repeated sampling of just single strain.

The initial study applying WGS to clinical Pf isolates collected from eight countries on three continents shows that the parasite exhibits significant population structure at continental scales, with the amount of subpopulation structure varying significantly among regions [23]. Employing novel F-statistics to measure the inbreeding coefficient, this work also argued that the degree of mixture varies significantly across populations, with highly mixed samples occurring relatively frequently in west Africa but only occasionally in Papua New Guinea. Importantly, the authors suggested an association between increased levels of observed mixture with increases in transmission intensity in the local environment. Transmission intensity, the rate at which individuals are infected with Pf, likely determines some part the frequency of out-crossing within parasite populations and so may be critical to understanding gene flow and strategies for resistance control [15].

In this paper, we present a new statistically rigorous model that synthesizes these two distinct and previously disparate approaches to analyzing *P. falciparum* clinical mixtures: assessing the number of distinct genetic types within a sample (the MOI approach) and measuring the degree of panmixia with respect to the local population (the inbreeding coefficient approach). The model centers around how these two sub-models contribute to generate the observed within-sample non-reference allele frequency as it relates to the population-level non-reference allele frequency for single nucleotide polymorphisms (SNPs). For clarity, we will deprecate the use of *non-reference* in front of the term allele frequency, since they are all calibrated in this fashion. We will use the acronyms WSAF to denote the within-sample allele frequency and PLAF to denote population-level allele frequency to avoid confusion about the particularly frequency being indicated.

The essential structure of the model is to explain observed 'bands' that emerge when examining the WSAF for SNPs as a function of their PLAF (Figure 1). The model posits that the number of these bands results as a direct consequence of the number of distinct strains present within a sample and that the degree of admixture with the local population determines bands' slopes. To distinguish from the inbreeding coefficient, we refer to the degree of admixture as the panmixia coefficient. The collection of bands are then modeled jointly as a finite mixture.

Figure 2 lays out the important components of the model. In the simplest case a sample is composed of a single, unmixed strain, and all SNPs exhibit a WSAF of zero or one (see Figure 2(a)), depending on whether they agree with the reference. Consequently the WSAF is independent of PLAF, leading to two flat bands at these values. We call these samples unmixed. In the case where finite number of strains mixed within a sample, then for each variant position some number of those strains will possess a reference allele and some will not. Which strains carry non-reference alleles and those strains' proportion in the sample mixture then determine the WSAF for each SNP. Observed across many SNPs, this leads to the apparent bands of constant WSAF across the PLAF. It follows that for K component strains there are 2^K possible combinations of biallelic states, leading to that number of apparent WSAF bands.

The complementary banding structure of panmixia arises when a fraction of the Pf organisms present within the blood are randomly sampled from the local population. In its simplest formulation, the panmixture model represents the admixture of two distinct Pf populations: a single strain, representing $1 - \alpha$ of the within-sample genomes, and a random sample of strains from the local population, representing α of the remaining genomes. In the case of perfect panmixia ($\alpha = 1$), a sample would be comprised of organisms evenly sampled from the ambient population and the plot of WSAF against PLAF would become a single line at $y = x$. In this data set, we do not observe any sample close perfect panmixia but observe several instances of apparent partial panmixia with a single dominant strain (Figure 1(c) and 2(c)). The α tilt in the WSAF arises from the fact that for this proportion of organisms the probability of sampling non-reference allelele is proportional to the PLAF, absent any other population structure. These samples, with a single strain and a degree of panmixia, we call panmixed. In more complex cases, where there is more than one dominant strain, the total number of bands is still determined by the number of these strains. However, now panmixia tilts each of these bands equally leading to complex mixtures (Figures 1(d) and 2(d)). The paper proceeds as follows. We first detail the structure of the WGS data, introduce some notation, and the essential mathematical structure of the model. We then present an extensive simulation study on the performance of the model and then an examination of its application to field isolates collected from northern Ghana. We conclude by discussing the strengths and weakness of the model, some possible improvements, and what consequences this analysis may have for understanding the etiology of clinical malaria.

DATA, NOTATION, AND MODEL

Data

The WGS data come from Illumina HiSeq sequencing applied to *P. falciparum* extracted from 235 clinical blood samples collected from infected patients from the Kassena-Nankana district (KND) region of Upper East region of northern Ghana. Collection occurred over approximately 2 years, from June 2009 to June 2011. The full sequencing protocol and collection regime are described in [23]. After quality control measures, sequencing was performed on 235 samples, and, following a documented protocol using comparison against world-wide variation, $198,181$ single-nucleotide polymorphisms (SNPs) were called within each sample [23]. Each call provides the number of reference and non-reference read counts observed at each variant position within the genome, ascertained against the the 3D7 reference [10]. For this project, we additionally filtered these data. First, multiallelic positions were reclassed as biallelic. We then excluded positions that exhibited no variation within the KND samples, any level of missingness (no read counts observed), or minor allele frequency less than 0.01. To remove low quality samples, we removed thoses possesed more than $4,000$ SNPs called with fewer than 20 read counts, following an inflection point observed in Supplementary Figure S1(a). These cleaning measures left $2,429$ SNPs in 168 samples. More than 95% of remaining samples' sequencing was completed without PCR amplification. We observe little apparent population structure among the samples, evidenced either by principal components analysis or a neighbor-joining tree of the pairwise difference among samples (Supplementary Figures S2). The data preparation scripts are available with the source code for the model, https://github.com/jacobian1980/pfmix/.

Notation

Following the data preparation and cleaning, our analysis begins with a set of N clinical samples, each composed of M SNPs. At each SNP j within each clinical sample i, we observe r_{ij} reads that agree with the reference genome and n_{ij} reads that do not agree with the reference. For a sample j. we write the complete data across all SNPs as $\mathcal{D}_i = [(r_{i1}, n_{i1}), \cdots, (r_{iM}, n_{iM})]$. For each SNP j, we associate a PLAF p_j. The collection of all p_j we refer to as $\mathcal{P}$.

Conditional upon the number of strains K, there are 2^K bands, indexed by $r = 1, \cdots, 2^K$. The full collection of bands we call $\mathcal{Q}$, with q_{ijr} indicating the WSAF for band r at SNP j in sample i. The probability of a SNP lying within the distinct bands across the PLAF is specified by a mixture component λ_r, which is a function of the PLAF, and so is often written $\lambda_r(p_j)$. The degree of panmixia in a sample is given by α, a value between zero and one. A complete list of the model parameters is given in Table 1.

Model

Statistically, the model takes the form of a finite mixture model, with the mixture components associated with individual bands [36, 26]. We take a Bayesian approach to inference and so lay out the model by giving an overall rationale for the decomposition of the posterior distribution and then justifying the appropriate choice of probability distributions for each of the terms [11].

Decomposition

We assume that samples are independent of each other and that the SNP data for each sample depends solely on K, the WSAF $\mathcal{Q}$, the PLAF $\mathcal{P}$, and a shape parameter ν. As samples are independent, we will deprecate sample-specific subscripts for the model parameters. Considering the data for a single sample, $\mathcal{D}_i$, the posterior distribution can then be written as:

$$\begin{aligned} \mathbb{P}(\mathcal{Q}, \mathcal{P}, \mathcal{W}, \alpha, \nu, K | \mathcal{D}_i) &\propto \mathbb{P}(\mathcal{D}_i | \mathcal{Q}, \mathcal{P}, \mathcal{W}, \alpha, \nu, K) \cdot \mathbb{P}(\mathcal{Q}, \mathcal{P}, \mathcal{W}, \alpha, \nu, K) \qquad (1) \\ &= \mathbb{P}(\mathcal{D}_i | \mathcal{Q}, \mathcal{P}, \nu, K) \cdot \mathbb{P}(\mathcal{Q}, \mathcal{P}, \nu, K, \mathcal{W}, \alpha). \qquad (2) \end{aligned}$$

We also assume that the WSAF, $\mathcal{Q}$, depends only on the PLAF, $\mathcal{P}$, the panmixia coefficient α, the number of strains K, and their proportions within the sample, $\mathcal{W}$, allowing the right-hand side of Equation 2 to be further decomposed, by noting that

$$\mathbb{P}(\mathcal{Q}, \mathcal{P}, \nu, K, \mathcal{W}, \alpha) = \mathbb{P}(\mathcal{Q} | \mathcal{P}, \nu, K, \mathcal{W}, \alpha) \cdot \mathbb{P}(\mathcal{P}, \nu, K, \mathcal{W}, \alpha). \qquad (3)$$

While $\mathcal{W}$ clearly depends on the number of strains, K, the remaining parameters we take to be independent of this value and of each other. This means that the last right-hand side term in Equation 3 becomes:

$$\mathbb{P}(\mathcal{P}, \nu, K, \mathcal{W}, \alpha) = \mathbb{P}(\mathcal{P}) \cdot \mathbb{P}(\nu) \cdot \mathbb{P}(\mathcal{W} | K) \cdot \mathbb{P}(K) \cdot \mathbb{P}(\alpha). \qquad (4)$$

Substituting Equations 3 and 4 into Equation 2, yields the final decomposition:

$$\begin{aligned} \mathbb{P}(\mathcal{Q}, \mathcal{P}, \mathcal{W}, \alpha, \nu, K | \mathcal{D}_i) \propto\ & \mathbb{P}(\mathcal{D}_i | \mathcal{Q}, \mathcal{P}, \nu, K) \cdot \mathbb{P}(\mathcal{Q} | \mathcal{P}, \nu, K, \mathcal{W}, \alpha) \cdot \\ & \mathbb{P}(\mathcal{P}) \cdot \mathbb{P}(\nu) \cdot \mathbb{P}(\mathcal{W} | K) \cdot \mathbb{P}(K) \cdot \mathbb{P}(\alpha). \qquad (5) \end{aligned}$$

We now specify each of the terms on the right-hand side above as probability distributions.

Likelihood : $\mathbb{P}(\mathcal{D}_i|\mathcal{Q}, \mathcal{P}, \nu, K)$

Within band r, the WSAF at SNP j in sample i is q_{ijr}. Supposing that read counts at j are identically and independently distributed with probability q_{ijr}, we model the probability of the data (r_{ij}, n_{ij}) as a Beta-binomial distribution, allowing us to model greater dispersion than expected under a pure binomial. We parameterize this distribution in terms of q_{ijr} and ν rather than the more commonly used shape and scale parameters, α and β. The relationship between the two parameterization is $q_{ijr} \cdot \nu = \alpha$ and $(1 - q_{ijr}) \cdot \nu = \beta$. We use this parameterization as it allows us to write the model in terms of the mean allele frequency that defines each band. The additional ν is a shape parameter that serves as a proxy for the variance. These parameters give a likelihood expression:

$$\mathbb{P}(n_{ij}, r_{ij}|r, q_{ijr}, \nu) \quad = \quad \binom{n_{ij} + r_{ij}}{n_{ij}} \cdot \frac{\mathrm{B}(n_{ij} + q_{ijr} \cdot \nu, r_{ij} + (1 - q_{ijr}) \cdot \nu)}{\mathrm{B}(q_{ijr} \cdot \nu, (1 - q_{ijr}) \cdot \nu)}, \tag{6}$$

where B is the beta function.

As any SNP could lie within any band, we employ a novel version of the finite mixture model to capture this segregation. Fixing the number of strains to K, there are then 2^K ways that the strains can be assorted into non-reference and reference allele states at any given position j. A given band r arises from C_r strains exhibiting the non-reference allele and $2^K - C_r$ strains exhibiting the reference allele. Supposing no population structure among the strains, the probability that a given SNP will be in that band is simply the probability of drawing C_r non-reference alleles and $2^K - C_r$ reference alleles, conditional upon p_j:

$$\begin{aligned} \mathbb{P}(\text{SNP } j \in \text{band } r|p_j) &= p_j^{C_r} \cdot (1 - p_j)^{2^K - C_r} \\ &= \lambda_r(p_j). \end{aligned}$$

Consequently, the density of the mixture coefficients for each band varies across the PLAF but such that they sum to 1 across all bands at any position j:

$$\begin{aligned} \mathbb{P}(\mathcal{D}_{ij}|\mathcal{Q}, \mathcal{P}, \nu, K) &= \sum_{r=1}^{2^K} \mathbb{P}(\text{SNP } j \in \text{band } r|p_j) \cdot \mathbb{P}(n_{ij}, r_{ij}|r, q_{ijr}, \nu) \\ &= \sum_{r=1}^{2^K} \lambda_r(p_j) \cdot \mathbb{P}(n_{ij}, r_{ij}|r, q_{ijr}, \nu). \end{aligned}$$

Following from the assumption of no population structure, SNPs will assort into bands independently. This leads to a product form for the likelihood of sample's data, $\mathcal{D}_i$:

$$\mathbb{P}(\mathcal{D}_i|\mathcal{Q}, \mathcal{P}, \nu, K) \quad = \quad \prod_{j=1}^{M} \left[\sum_{r=1}^{2^K} \lambda_r(p_j) \cdot \mathbb{P}(n_{ij}, r_{ij}|r, q_{ijr}, \nu) \right]. \tag{7}$$

Band structure: $\mathbb{P}(\mathcal{Q}|\mathcal{P}, \nu, K, \mathcal{W}, \alpha)$

The full mixture model contains two distinct subcomponents that we call the simple mixture model and the panmixture model, respectively. Both models generalize the unmixed case, though naturally in different ways. We first describe the unmixed model and then layout the two extensions before showing how these can be combined to create the full model. In practice, we only fit data using the full model and allow it to indicate the number of strains, their proportions, and the degree of panmixia. We do not know the number of strains *a priori* so we employ metrics applied to the posterior distribution inferred with different values of K to determine it. However, for the purpose of detailing the model, we assume that K is known.

Unmixed model - In an unmixed sample only one strain is present and there is no panmixia, and so $K = 1$ and $\alpha = 0$. Consequently, we expect all SNPs to exhibit WSAF either zero or one (Figure 2(a)) depending on whether they agree with the reference or not. There are then two bands, $r = 1, 2$ and $q_{ij1} = 0$ and $q_{ij2} = 1$.

Simple mixture model - The simple mixture model assumes that a finite number K of distinct strains, $s_1, \cdots, s_K$, are combined together in the sample with proportions, $\mathcal{W} = [w_1, \cdots, w_K]$ but that $\alpha = 0$. Naturally, $\sum_k w_k = 1$ and for each SNP j, the probability of being within band r is given by $\lambda_r(p_j)$, as above. Band r is defined by a vector $v_r = [\mathbf{1}_{\{s_1 \in r\}}, \cdots, \mathbf{1}_{\{s_K \in r\}}]$, where $\mathbf{1}_{\{s_k \in r\}}$ is an indicator function of whether strain k exhibits a non-reference allele within the sample. The WSAF of q_{ijr} is then given by the sum of all of proportions of strains that exhibit a non-reference allele:

$$q_{ijr} = \sum_{k=1}^{K} w_k \cdot \mathbf{1}_{\{s_k \in r\}}. \tag{8}$$

Taken across all r bands, this leads to 2^K bands with zero slope and corresponding proportions $(0, w_1, \cdots, w_K, w_1 + w_2, w_1 + w_3, \cdots, 1)$.

Panmixture model - As mentioned above, in its simplest case, the panmixture model represents the admixture of two distinct Pf populations: a single strain, representing $1 - \alpha$ of the within-sample orgnaisms, and a random sample of strains from the local population, for the remaining α organisms. When $\alpha = 0$ the model reduces to the unmixed case. We will refer the single strain as the dominant strain, although, conditional upon α, it may represent only a small proportion of the sample's population. For each position j, there are still only two bands: the higher one corresponding to the non-reference allele being present in the dominant strain, and the lower one corresponding to its absence. However, the WSAF for these bands varies according to p_j with slope α. To resolve q_{ijr}, first consider the upper band, $r = 2$. At any position j, $1 - \alpha$ of the reads come from the dominant strain. The remaining reads, each sampled randomly from the local population, each have probability p_j of being non-reference. This leads to $q_{ij2} = (1 - \alpha) + \alpha \cdot p_j$. For the lower band, the dominant strain contributes no non-reference reads so $q_{ij1} = \alpha \cdot p_j$.

Complex mixture model - The complex model synthesizes the simple mixture and panmixture models. In this case, at position j, α of the reads are sampled randomly from the across the local population, contributing a fraction of $\alpha \cdot p_j$ non-reference alleles. The state of the remaining reads are determined by $\mathcal{W}$ as in Equation 8. For band r at position j, the WSAF is then given by

$$q_{ijr} = (1 - \alpha) \cdot \left(\sum_{k=1}^{K} w_k \cdot \mathbf{1}_{\{s_k \in r\}} \right) + \alpha \cdot p_j. \tag{9}$$

There are then 2^K bands with proportions $(0, w_1, \cdots, w_K, w_1 + w_2, w_1 + w_3, \cdots, 1)$ and slope α.

Priors

For the remaining four probability distributions we place the following vague prior distributions:

$$\begin{aligned}
\mathcal{W}|K &\sim \text{DIRICHLET}(\mathbf{1}_K) \\
\alpha &\sim \text{UNIFORM}(0,1) \\
\nu &\sim \text{EXPONENTIAL}(5) \\
K &\sim \text{zero-truncated POISSON}(2),
\end{aligned}$$

where $\mathbf{1}_K$ is a vector of K ones.

Inference

We infer the model parameters using a standard Bayesian Markov chain Monte Carlo (MCMC) approach [13, 12] with one exception: we first calculate maximum-likelihood estimates (MLE) for $\mathcal{P}$ across all samples and then treat these as fixed when inferring the remaining parameters [38]. This choice is motivated by statistical expedience and computational speed. Except for $\mathcal{P}$, the parameters of the model are independent across samples and so this approximation enables the algorithm to infer parameters in parallel rather than jointly. This avoids the difficulties of performing inference on the number of strains within samples simultaneously, which would require an involved trans-dimensional MCMC scheme (such as reversible jump MCMC, [14]) acting jointly across all samples. Running in parallel also increases the computational speed of the implementation by at least an order of magnitude. Since the sample collection is large enough that $\mathcal{P}$ is nearly independent of any given sample, we do not expect this approximation to significantly bias inference.

For each SNP j, the MLE derives from treating the non- and reference reads within a sample as coming from a binomial distribution with parameter p_j. This leads to:

$$\hat{p}_j \quad = \quad \sum_i^N n_{ij} \Big/ \sum_i^N (n_{ij} + r_{ij}).$$

To infer K for each sample, we employ a Bayesian Information Criterion (BIC) [35, 6] and harmonic mean estimator to the Bayes Factor (hmeBF) [22, 20] as metrics for model selection. To find the maximum likelihood value for use with the BIC, we initially implemented a separate estimation algorithm but found no significant difference with using the highest value observed from the posterior samples. In simulations, we observe that the BIC and hmeBF provide similar guidance, with the BIC frequently indicating a smaller K. For the simulation study and empirical data example, we provide only the BIC result.

Conditional on $\mathcal{P}$ and K, we implement a Metropolis-Hastings algorithm to draw samples from the posterior distribution [13]. For each of the three parameters, α, $\mathcal{W}$, and ν, we propose new values directly from the prior distribution, leading to Metropolis-Hastings ratios almost solely dependent on the ratio between the likelihood and priors for the proposed state to those for the current. The inference scheme is implemented in set of scripts for the R computing language, and can be found under the Academic Free License at https://github.com/jacobian1980/pfmix/s. For a single sample, a sufficiently long MCMC run takes approximately 20 minutes on a single high-performance computing core.

RESULTS

Simulations under the model

To demonstrate the efficacy of our implementation, we present a simulation study examining the algorithm's performance. We consider two distinct aspects of the inference separately: how well the model infers the number of strains, and, conditional upon that number, how well it infers the model's other parameters. We simulate data from the model in the following way. Conditional upon M,α, K and the sum of the read counts, C, we draw a vector of probabilities, $\mathcal{W}$, from a uniform Dirichlet distribution. We combine the values of $\mathcal{W}$ in all possible permutations to create the 2^K bands and assign the PLAF for the SNPs evenly from $1/M$ to 1, so that the j^{th} SNP has PLAF $\frac{j}{M}$. For each SNP, we first probabilistically select the band it occupies according to according to Equation 7. Conditional upon selecting r, we then simulate read counts according to the likelihood (Equation 6) with q_{ijr} according to Equation 9. For all simulations, we set $\nu = 10$. We run the simulation across the range of values for M,α, K and C. For each parameter set, we create 10 independent realizations.

Number of components

Figure 3 shows performance of the algorithm for inferring the number of components increases in precision with the number of SNPs and the number of reads. Conditional upon α, the simulations indicate that the number of SNPs, M, to be the largest determinant of performance, with the sum of the read counts, C, playing an important supporting role. Inference of the number of underlying strains, K, is generally strong for low panmixture levels (small α values), but is noticeably more conservative for $\alpha = 0.5$, likely due to the bands becoming increasingly tightly packed as panmixia increases. In general, inference is slightly conservative, likely owing to the BIC estimator's bias toward parsimony [9].

Parameters

Figure 4 shows similar performance for inference of the strain proportions $\mathcal{W}$ and α. For $\mathcal{W}$, we report the mean squared deviation. For α, we report the absolute normalized deviation to account for relative difference from the true value. For both parameters, we observe that the number of SNPs is the strongest determinant of accuracy, with $M = 150$ ensuring moderately strong performance. High α moderately decreases the quality of inference for the strain proportions.

Clinical samples from northern Ghana

Applying the algorithm to the 168 high-quality samples from KND, we observe K range 1 to 7, with α falling between 0 and 0.14, and a moderate correlation between K and α (Figure 5). The largest subset of samples were unmixed, with $K = 1$ and $\alpha < 0.01$, though the majority of samples exhibit moderate levels of mixture, with $K = 2, 3, 4$ and $0.01 \leq \alpha \leq 0.03$. A small number of samples exhibit complex mixtures, with $K > 4$ and α typically greater than 0.02. These results confirm the presence of mixtures within Pf clinical isolates, but also indicate, possibly more complex patterns involving interactions between the number of dominant strains and the degree of panmixia.

We observe that for most samples the 95% credible interval for α is within a small percentage of the median value. For $\mathcal{W}_i$, we observe a similarly tight posterior distribution, particularly for samples with $K \leq 3$. The posterior uncertainty increases together with increasing K and increasing α. For a small number of samples, the model initially produced unusually low values for ν, indicating a bimodal Beta-binomial distribution inconsistent with a mixture of strains and consequently suspect values for K. For these, we bounded ν such that it ensured a unimodal distribution and then recovered results consistent the remaining samples.

To visually inspect the quality of the results, we generate figures for each of the samples showing the observed WSAF and PLAF data, the inferred model structure, and data simulated under the inferred model following the observed PLAF. We show examples of these plots for three typical samples in Figures 6. Nearly all samples (158/168), across all different mixture patterns, show strong visual correspondence between the observed and model-simulated data. We also observe a strong correlation between the inferred number of components and a quasi-maximum likelihood estimate for the inbreeding coefficient for each sample (Figure 7) [31].

For each sample, we compare the full model to two reduced versions under the restrictions $\alpha = 0$ and $K = 1$, respectively. These restrictions correspond to the cases where the model becomes the simple mixture model, with 2^K bands but no admixture with the local population, and the panmixture model, where there is a single strain with some local population admixture, respectively. For numerical stability reasons, we set the former restriction as $\alpha = 0.001$. For 60% of samples (108/168), the BIC selects the full model over either of the restricted models. For 58 samples the BIC criterion selected the $K = 1$ restriction, while for 23 samples it selected $\alpha = 0$. Taken in aggregate across all samples, the criterion overwhelmingly selects the full model over either of the restricted models.

DISCUSSION

The model captures two dimensions of within-sample mixture for *P. falciparum* that had previously modeled separately: the number of strains and the degree of admixture with the local population. Evidenced by the comparison of the full model with restricted sub-models, this approaches provides a marked improvement over both more restricted approaches in capturing the structure of mixture in clinical samples. While the model provides a more involved qualitative understanding of the samples, the strong correlation between the inbreeding coefficient and the inferred number of strains shows that the model produces results consistent with previous methods.

In order to perform inference, the model makes a number of simplifying assumptions that may be violated in practice. The model presumes that SNPs are unlinked and consequently independent for the purpose of calculating the likelihood. Given the high recombination rate of *P. falciparum* this assumption may hold for the majority of pairs of SNPs, but neglects correlations that appear locally ($\sim$ 10 kB). However, we expect that this independence assumption serves to moderately weaken the inferential power of the model rather than cause any type of bias since it fails to include possibly informative data, rather than posit a possibly misspecified model. More problematic is the model's implicit assumption of limited population structure. In the case of the KND samples, and perhaps in much of west Africa, this assumption appears supported [1, 23]. In other contexts, specifically south-east Asia, recent population bottlenecks and selection suggest that this assumption will be violated [27]. The consequences on this model inference are unknown but can likely be partially resolved with appropriate simulation studies.

The model presents an important new tool for interrogating the biology of clinical Pf infections. In particular, how the number of component strains and the panmixia coefficient relate to the infection parameters, such as seasonality, transmission intensity, and outcrossing, and evolutionary parameters such as the rate of change within sections of the Pf genome, could have implications for understanding the genetic epidemiology of Pf. The model also presents a means for clarifying the poorly detailed structure of intra-host infection dyanmics, such as strain selection or density-dependent selection [21], by resolving how the number of strains, the mixture proportions, and the panmixia coefficient change within the course of an infection or in response to drug intervention.

An unexpected structural consequence of the model is that power to infer additional strains diminishes as the panmixia coefficient (α) increases. This results from the simplifying assumption that $1 - \alpha$ of the reads come from the dominant strains while the remaining α fraction are sampled randomly from the local population. Geometrically, we see that as α increases that the bands will get progressively closer together as they approach panmixia, making them harder for the model to distinguish. Also, as α increases to one, the fraction of reads representing the dominant strains diminishes, reducing power to infer these components. We observe this pattern strongly in simulations (Figure 4): for $\alpha = 0.5$ or greater, the model consistently infers too few components. While this deficiency may be overcome in some fashion with additional SNPs or read counts, the geometry indicates there may be fundamental limits on any model's ability to discriminate the true number of components in the high panmixia regime.

In principle, the model can be explicitly tested against experiment. Laboratory facilities with the capacity to store many field strains (> 100) could generate artificial samples in an experimental analog of our simulation procedure, as follows. Starting with N unmixed strains, identified using inbreeding coefficients, they could create mixtures, they would need to first fix the required sequencing volume as η, and the parameters for panmixia (α), number of component strains (K), and their mixture parameters, $\mathcal{S}$ and $\mathcal{W}$. For the finite mixture component, they would then combine volumes of $\eta \cdot \mathcal{W}$ from the K strains. For the panmixture component, they would then fix some large number but experimentally feasible number of strains (say 100) and randomly sample from all of them a volume of $\eta/100$. Finally, combining these into final sample and applying WGS sequencing, will yield data that we hypothesize will closely follow the integrated model outlined above, with ν capturing the experimental variation. Naturally, consistent results would indicate the sufficiency of the model, but not it's necessity, holding out the possibility of a more minimal description. These results could be further compared against other next-generation technologies, such as single-cell sequencing, that have been deployed to understand Pf clinical mixtures [30].

The method works efficiently in practice (the properties of a single sample can be inferred in minutes on a standard laptop) but a number of possible improvements could strengthen its statistical performance. Most immediately, creating a full Bayesian approach rather than the parallelizing implementation here - while likely not improving the parametric inference for individual samples - would provide the full posterior distribution across all samples for more considered model comparison. In that same line, more refined approaches to inferring the number of strains within samples, either via a reversible jump MCMC approach or methods for rigorously estimating Bayes factors [14], would provide researchers more power to resolve the structure of mixture, though likely at the cost of significantly more computation.

The model does not perform haplotype phasing to resolve the sequence of the underlying strains [43, 18, 32]. The analysis here suggests that a method for estimating haplotypes would be straightforward for some samples (unmixed ones, for instance) but difficult if not impossible for others (when α is greater than 0.5). Researchers may be particularly interested in whether, in these phased samples, particular stretches of the genome appear more or less frequently in the dominant strains than others, indicating immunological or environmental selection. This is also an avenue for statistical development.

The two phenomenologies of mixture that the model captures - a finite mixture of distinct strains and an inbred population admixture - cannot be immediately associated with any specific aspect of the infection process. A number of variables appear plausible in determining these relationships, including transmission intensity, the length of infection, the immunological status of the infected individual, and within-host density dependent selection. Together with WGS data, this new approach can serve as a means for biological researchers to directly resolve these hypotheses and resolve the consequence of mixture in *P. falciparum* infections.

Authorship

JO designed and implemented the study and wrote the manuscript. ZI assisted in the design of the study and commented on the manuscript. LA-E collected the data and commented on the manuscript and the study.

Tables

Parameter	Definition
N	Number of samples
M	Number of SNPs
K	Number of strains
$i = 1, \cdots, N$	Index for samples
$j = 1, \cdots, M$	Index for SNPs
$r = 1, \cdots 2^K$	Index for bands / strain mixtures
p_j	(Non-reference) allele frequency for SNP j
$\mathcal{P} = [p_j]$	The PLAF for all SNPs
$\mathcal{Q} = [q_{ij}]$	Within-sample allele frequency for SNP j in sample i
α	Degree of panmixia within a sample, panmixia coefficient
$\mathcal{S} = [s_1, \cdots, s_K]$	Strains in a sample
$\mathcal{W} = [w_1, \cdots, w_K]$	Strain proportions in a sample
λ_r	Band proportions within sample
ν	Variation parameter for Beta-binomial
WSAF	Within-sample allele frequency
PLAF	Population-level allele frequency

Table 1: Parameters and definitions for the model and its description.

Parameter	Values:			
M	50	150	500	2500
C	10	25	100	250
α	0.01	0.1	0.5	
K	1	3		

Table 2: Table of simulated parameter values: C the number of read counts while M, K and α are as in Table 1.

Figures

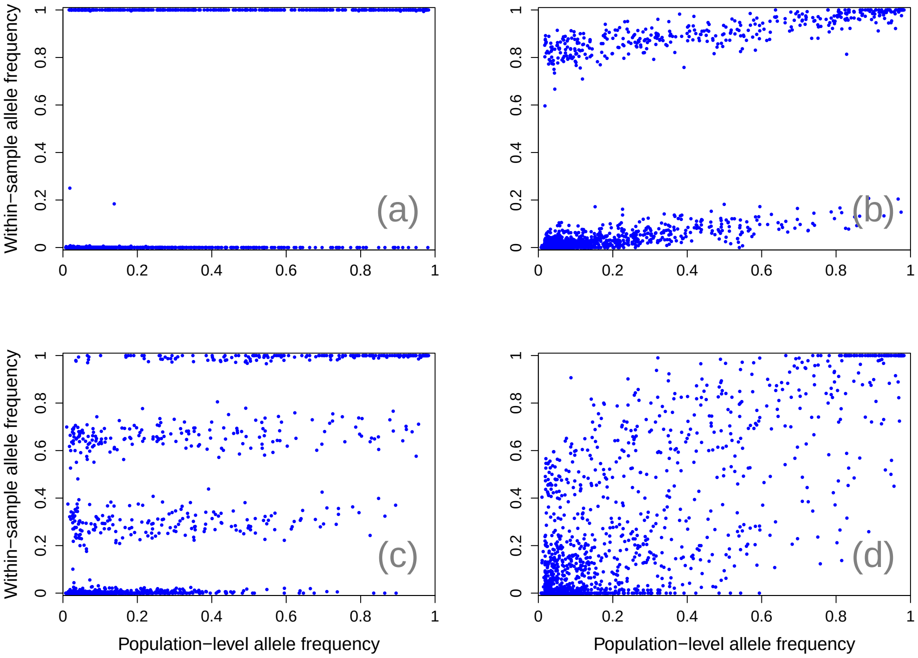


Figure 1: Four representative samples with WSAF for each SNP plotted against the PLAF, showing an absence of mixture (a), a partially panmixed sample (b), a simple mixture (c), and a complex mixture (d).

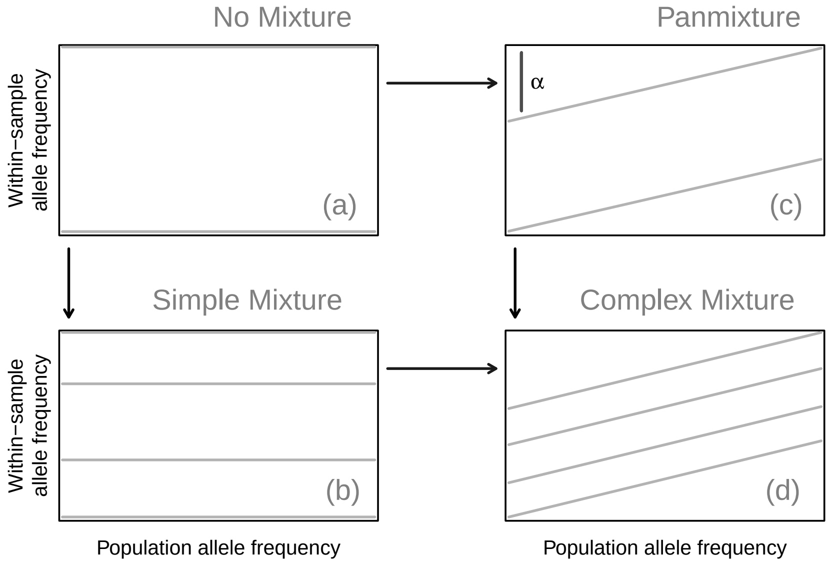

No Mixture
Panmixture
α
Within−sample allele frequency
(a)
(c)
Simple Mixture
Complex Mixture
Within−sample allele frequency
(b)
(d)
Population allele frequency
Population allele frequency

Figure 2: The essential structure of the model comprises four distinct states, relating the WSAF to the PLAF: no mixture (upper left); simple mixture (lower left); panmixture (upper right); and complex mixture (lower right).

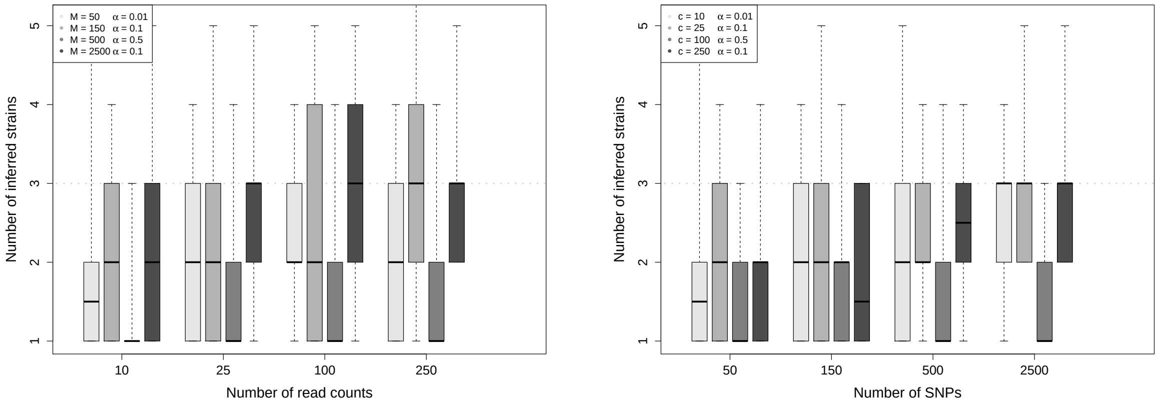
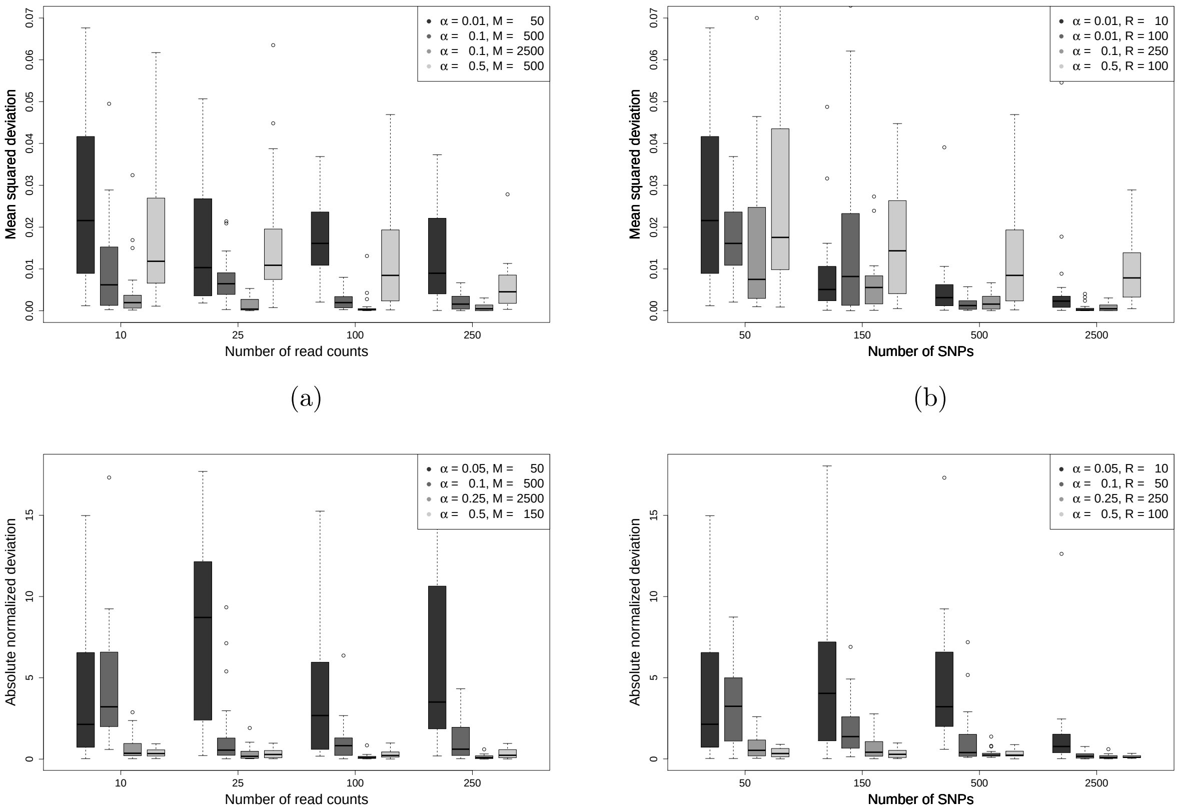


Figure 3: Performance for inference of number of components



(a)

(b)

(c)

(d)

Figure 4: Performance for parameter inference: (a) mean squared deviation for $\mathcal{W}$ by number of read counts; (b) mean squared deviation for $\mathcal{W}$ by number of SNPs; (c) absolute normalized deviation for α by number of read counts; and (d) absolute normalized deviation for α by number of SNPs.

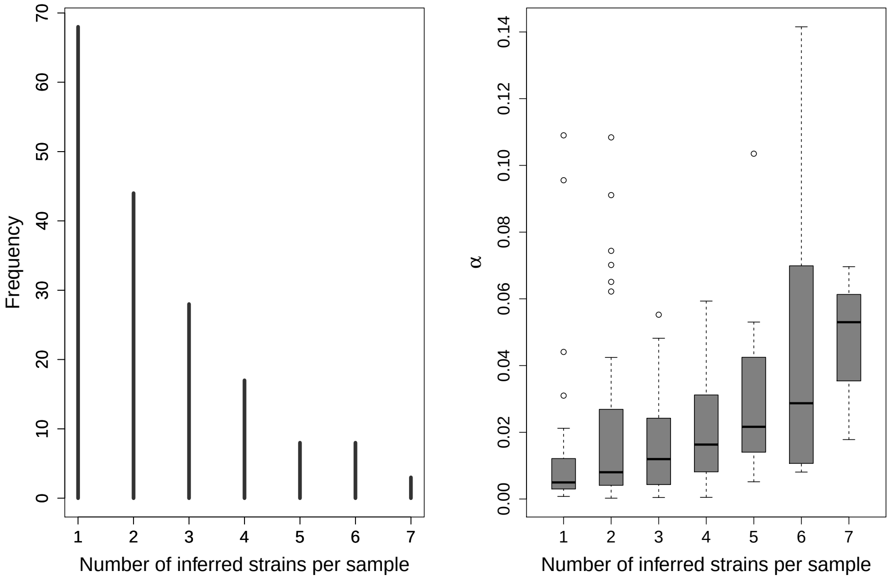


Figure 5: The frequency of number of inferred strains per sample (lef) and the posterior median value of α by the number of inferred strains (right).

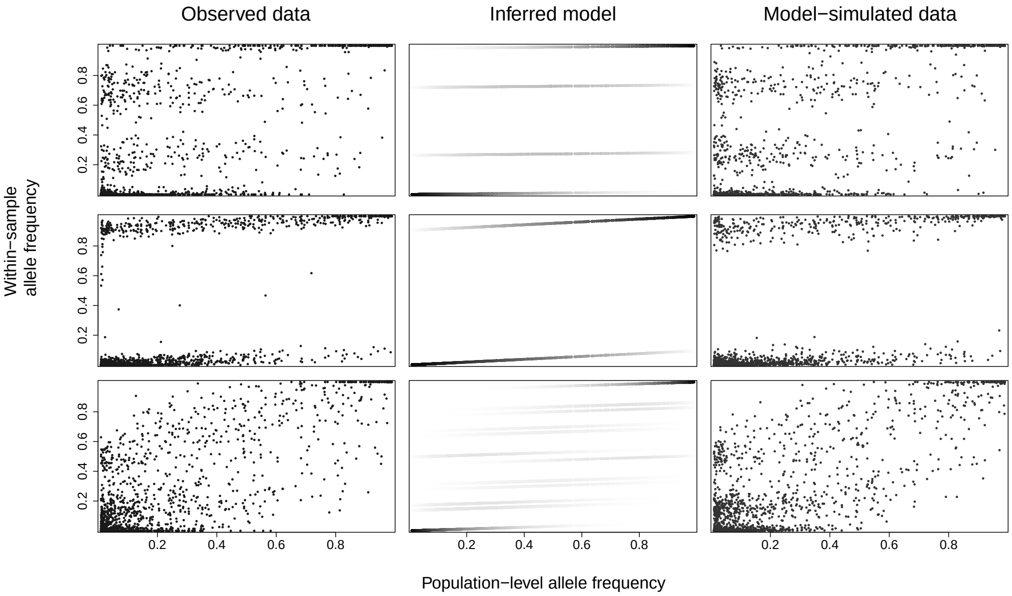


Figure 6: Examples of real data. For three samples (rows), we present the observed data WSAF plotted against the PLAF (first column), a diagram of the inferred model indicating the bands, proportions, and α (second column), and data simulated under the inferred model. α and $\mathcal{W}$ are the maximum *a posteriori* values. In the second column, the model's PLAF-varying mixture densities are shown in grey scale, with black equal to one.

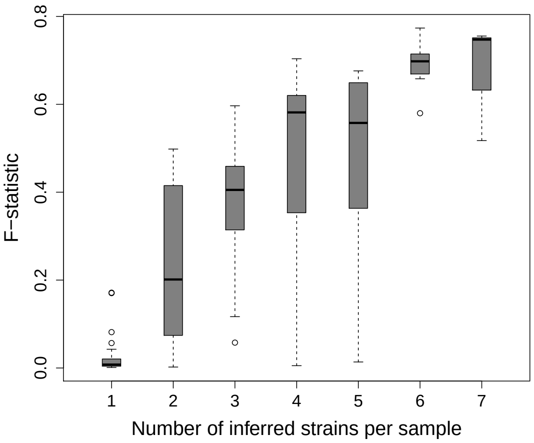


Figure 7: Boxplot of the F-statistic inbreeding coefficient $(1 - F_{is})$ for each sample grouped by the number of inferred strains.

References

[1] Timothy JC Anderson, Bernhard Haubold, Jeff T Williams, Jose G Estrada-Franco, Lynne Richardson, Rene Mollinedo, Moses Bockarie, John Mokili, Sungano Mharakurwa, Neil French, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Molecular Biology and Evolution*, 17(10):1467–1482, 2000.

[2] Sarah Auburn, Susana Campino, Taane G Clark, Abdoulaye A Djimde, Issaka Zongo, Robert Pinches, Magnus Manske, Valentina Mangano, Daniel Alcock, Elisa Anastasi, et al. An effective method to purify *Plasmodium falciparum* dna directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS ONE*, 6(7):e22213, 2011.

[3] Sarah Auburn, Susana Campino, Olivo Miotto, Abdoulaye A Djimde, Issaka Zongo, Magnus Manske, Gareth Maslen, Valentina Mangano, Daniel Alcock, Bronwyn MacInnis, et al. Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PloS one*, 7(2):e32891, 2012.

[4] H-P Beck, I Felger, P Vounatsou, R Hirt, M Tanner, P Alonso, and C Menendez. 8. effect of iron supplementation and malaria prophylaxis in infants on *Plasmodium falciparum* genotypes and multiplicity of infection. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(Supplement 1):41–45, 1999.

[5] STEFANIE Beck, Frank P Mockenhaupt, Ulrich Bienzle, Teunis A Eggelte, WN Thompson, and Klaus Stark. Multiplicity of *Plasmodium falciparum* infection in pregnancy. *The American journal of tropical medicine and hygiene*, 65(5):631–636, 2001.

[6] Scott Shaobing Chen and Ponani S Gopalakrishnan. Clustering via the bayesian information criterion with applications in speech recognition. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 645–648. IEEE, 1998.

[7] DJ Conway, BM Greenwood, and JS McBride. The epidemiology of multiple-clone *Plasmodium falciparum* infections in Gambian patients. *Parasitology*, 103(Pt 1):1–6, 1991.

[8] Anna Färnert, Ingegerd Rooth, Åke Svensson, Georges Snounou, and Anders Björkman. Complexity of *Plasmodium falciparum* infections is consistent over time and protects against clinical disease in tanzanian children. *Journal of infectious diseases*, 179(4):989–995, 1999.

[9] David F Findley. Counterexamples to parsimony and BIC. *Annals of the Institute of Statistical Mathematics*, 43(3):505–514, 1991.

[10] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511, 2002.

[11] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[12] Charles J Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, pages 473–483, 1992.

[13] Walter R Gilks. *Markov chain Monte Carlo*. Wiley Online Library, 2005.

[14] Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[15] Carlos A Guerra, Priscilla W Gikandi, Andrew J Tatem, Abdisalan M Noor, Dave L Smith, Simon I Hay, and Robert W Snow. The limits and intensity of *Plasmodium falciparum* transmission: implications for malaria control and elimination worldwide. *PLoS medicine*, 5(2):e38, 2008.

[16] Simon I Hay, Carlos A Guerra, Peter W Gething, Anand P Patil, Andrew J Tatem, Abdisalan M Noor, Caroline W Kabaria, Bui H Manh, Iqbal RF Elyazar, Simon Brooker, et al. A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS Medicine*, 6(3):e1000048, 2009.

[17] L Henning, D Schellenberg, T Smith, D Henning, P Alonso, M Tanner, H Mshinda, H-P Beck, and I Felger. A prospective study of *Plasmodium falciparum* multiplicity of infection and morbidity in tanzanian children. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 98(12):687–694, 2004.

[18] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8):955–959, 2012.

[19] Ghazi A Jamjoom. Dark-field microscopy for detection of malaria in unstained blood films. *Journal of clinical microbiology*, 17(5):717–721, 1983.

[20] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[21] Dominic Kwiatkowski and Martin Nowak. Periodic and chaotic host-parasite interactions in human malaria. *Proceedings of the National Academy of Sciences*, 88(12):5111–5113, 1991.

[22] Michael Lavine and Mark J Schervish. Bayes factors: what they are and what they are not. *The American Statistician*, 53(2):119–122, 1999.

[23] Magnus Manske, Olivo Miotto, Susanna Campino, Sarah Auburn, Jacob Almagro-Garcia, Gareth Maslen, Jack O'Brien, and Dominic Kwiatkowski. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*, AOP, 2012.

[24] Pembe Issamou Mayengue, Adrian JF Luty, Christophe Rogier, Meili Baragatti, Peter G Kremsner, and Francine Ntoumi. The multiplicity of *Plasmodium falciparum* infections is associated with acquired immunity to asexual blood stage antigens. *Microbes and Infection*, 11(1):108–114, 2009.

[25] IA McGregor. Immunology of malarial infection and its possible consequences. *British medical bulletin*, 28(1):22–27, 1972.

[26] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[27] Olivo Miotto, Jacob Almagro-Garcia, Magnus Manske, Bronwyn MacInnis, Susana Campino, Kirk A Rockett, Chanaki Amaratunga, Pharath Lim, Seila Suon, Sokunthea Sreng, et al. Multiple populations of artemisinin-resistant *P* in cambodia. *Nature genetics*, 45(6):648–655, 2013.

[28] Toshihiro Mita, Kazuyuki Tanabe, and Kiyoshi Kita. Spread and evolution of *Plasmodium falciparum* drug resistance. *Parasitology international*, 58(3):201–209, 2009.

[29] DA Müller, JD Charlwood, I Felger, C Ferreira, V Do Rosario, and T Smith. Prospective risk of morbidity in relation to multiplicity of infection with *Plasmodium falciparum* in São Tomé. *Acta tropica*, 78(2):155–162, 2001.

[30] Shalini Nair, Standwell C Nkhoma, David Serre, Peter A Zimmerman, Karla Gorena, Benjamin J Daniel, François Nosten, Timothy JC Anderson, and Ian H Cheeseman. Single-cell genomics for dissection of complex malaria infections. *Genome research*, 24(6):1028–1038, 2014.

[31] John D O'Brien and Lucas Amenga-Etago. Approaches to estimatg inbreeding coefficients within clinical isolates of plasmodium falciparum from genomic sequence data. *under review*, ., 2014.

[32] John D O'Brien, Xavier Didelot, Zamin Iqbal, Lucas Amenga-Etego, Bartu Ahiska, and Daniel Falush. A Bayesian approach to inferring the phylogenetic structure of communities from metagenomic data. *Genetics*, pages 114–119, 2014.

[33] Giacomo M Paganotti, Hamza A Babiker, David Modiano, Bienvenu S Sirima, Federica Verra, Amadou Konate, Andre L Ouedraogo, Amidou Diarra, Margaret J Mackinnon, Mario Coluzzi, et al. Genetic complexity of *Plasmodium falciparum* in two ethnic groups of burkina faso with marked differences in susceptibility to malaria. *The American journal of tropical medicine and hygiene*, 71(2):173–178, 2004.

[34] D Payne. Spread of chloroquine resistance in *Plasmodium falciparum*. *Parasitology today*, 3(8):241–246, 1987.

[35] David Posada and Thomas R Buckley. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808, 2004.

[36] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

[37] Cally Roper, Richard Pearce, Shalini Nair, Brian Sharp, François Nosten, and Tim Anderson. Intercontinental spread of pyrimethamine-resistant malaria. *Science*, 305(5687):1124–1124, 2004.

[38] FW Scholz. Maximum likelihood estimation. *Encyclopedia of statistical sciences*, 1985.

[39] Amar Bir Singh Sidhu, Dominik Verdier-Pinard, and David A Fidock. Chloroquine resistance in *Plasmodium falciparum* malaria parasites conferred by pfcrt mutations. *Science*, 298(5591):210–213, 2002.

[40] T Smith, H-P Beck, A Kitua, S Mwankusye, I Felger, N Fraser-Hurt, A Irion, P Alonso, T Teuscher, and M Tanner. 4. Age dependence of the multiplicity of *Plasmodium falciparum* infections and of other malariological indices in an area of high endemicity. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(Supplement 1):15–20, 1999.

[41] T Smith, I Felger, N Fraser-Hurt, and H-P Beck. 10. Effect of insecticide-treated bed nets on the dynamics of multiple *Plasmodium falciparum* infections. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93(Supplement 1):53–57, 1999.

[42] Robert W Snow, Carlos A Guerra, Abdisalan M Noor, Hla Y Myint, and Simon I Hay. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature*, 434(7030):214–217, 2005.

[43] Matthew Stephens, Nicholas J Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, 2001.

[44] Heide A Stirnadel, Ingrid Felger, Tom Smith, Marcel Tanner, Hans-Peter Beck, et al. Malaria infection and morbidity in infants in relation to genetic polymorphisms in tanzania. *Tropical Medicine & International Health*, 4(3):187–193, 1999.

[45] Bruce S Weir and C Clark Cockerham. Estimating F-statistics for the analysis of population structure. *evolution*, pages 1358–1370, 1984.

[46] RJM Wilson, IA McGregor, K Williams, P Hall, and R Bartholomew. Antigens associated with *Plasmodium falciparum* infections in man. *The Lancet*, 294(7613):201–205, 1969.

[47] John C Wootton, Xiaorong Feng, Michael T Ferdig, Roland A Cooper, Jianbing Mu, Dror I Baruch, Alan J Magill, and Xin-zhuan Su. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature*, 418(6895):320–323, 2002.

[48] World Health Organization. *World malaria report 2008*. World Health Organization, 2008.